Strong Enforcement by a Weak Authority

Strong Enforcement by a Weak Authority*

Jakub Steiner†
CERGE-EI

February 17, 2006

Abstract

This paper studies the enforcement abilities of authorities with a limited commitment to punishing violators. Commitment of resources sufficient to punish only one agent is needed to enforce high compliance of an arbitrary number of agents. Though existence of other, non-compliance equilibria is generally inevitable, there exist punishment rules suitable for a limited authority to assure that compliance prevails in the long run under stochastic evolution.

JEL classification: C73, D64, H41.

Keywords: Commitment, Enforcement, Punishment, Stochastic Evolution.

*The paper builds on my earlier work "A Trace of Anger is Enough, on the Enforcement of Social Norms". I benefited from the comments of Kenneth Binmore, Fuhito Kojima, Simon Gächter, Werner Güth, Eugen Kováč, and Jaromír Kovařík. Dirk Engelmann, Andreas Ortmann, and Avner Shaked inspired me in numerous discussions. Laura Strakova carefully edited the paper. The usual disclaimer applies.

†Center for Economic Research and Graduate Education, Charles University, and Economics Institute, Academy of Sciences of the Czech Republic (CERGE-EI), Address: Politickych Veznu 7, 111 21, Prague, Czech Republic, Tel: +420-605-286-947, E-mail: jakub.steiner@cerge-ei.cz. WWW: http://home.cerge-ei.cz/steiner/

1 Introduction

Centralized authorities, such as governments, or decentralized ones, such as peers, use threats of punishment to enforce norms. However the authority, whether centralized or decentralized, achieves compliance only if it is able to commit to the punishment threat. Punishment is often costly, and hence an important determinant of the authority's success at enforcement is the amount of resources committed for punishment. In this paper I argue that both kinds of authorities are similar in that they can enforce high compliance of many agents with only few committed resources. The argument is as follows: suppose that the authority is *limited* in that it can commit only resources that suffice to punish just one agent by an amount higher than the agent's cost of compliance. Then, the authority's punishment threat induces among the agents a game with an equilibrium, in which all agents comply, as no agent wishes to deviate individually. For a centralized authority this implies that it is able to control an arbitrary number of subordinates as long as it is able to control one. Similarly, it is possible to apply this observation to decentralized peer enforcement in a public good game with punishment option. N players, each committing one unit for punishment, can enforce individual contributions of approximately N units, and to collect altogether approximately N^2 units.

However, even though a small punishment commitment may deter individual defectors from deviating, the existence of a non-compliance equilibrium would appear to be unavoidable. The committed resources are insufficient to punish all and therefore, if no agent comply, the punishment of each is small compared to the cost of compliance. Yet, as shown below, any limited authority may avoid the non-compliance equilibrium — at least in the long run — by choosing a proper punishment rule. The supporting argument is contingent on the authority's ability to punish colluding violators at least slightly. I divide authorities into two categories along this line. *Collusion-vulnerable* authorities cannot punish if all agents coordinate on the same level of non-compliance. Anger-based peer enforcement is a prime example, because punishing after a perfect collusion would require the punisher to be angry with peers who have perpetrated the same offense as herself.[1] *Collusion-resistant* authorities are able to punish by at least some amount even after a perfect collusion. The punishment of each colluding agent can be arbitrarily small so even an authority with limited committed resources can be

[1]Decentralized authority based on peer enforcement may more frequently belong to this category but even a centralized authority such as a government may be constrained, for instance politically, to punish agents unified in a common non-compliance action.

collusion-resistant.

Let us first present a punishment rule which eliminates the non-compliance equilibria yet which is suitable for a limited collusion-resistant authority. The rule requires the authority to commit to punishing only the worst violator. In the case of a tie the authority divides the punishment equally among the worst violators which in turn induces a dominance solvable game among the agents. The lowest possible compliance level is dominated as it guarantees punishment, and an increase in compliance just above the second lowest level saves the violator from punishment. Elimination continues by induction until only high compliance levels remain in the strategy sets.

A collusion-vulnerable authority, in contrast, cannot use the above "punish-the-worst" rule as it requires slight punishment of all players even after a perfect collusion. An equilibrium in which no players comply inevitably exists under a collusion-vulnerable authority as no player can be punished in such an equilibrium. To assess which equilibrium prevails in the long run, I build a stochastic, evolutionary model along the lines of Kandori, Mailath and Rob (1993). Agents occasionally but rarely deviate from their best responses and experiment with a random action. As demonstrated below, only a high level of compliance survives the evolution under a simple punishment rule.

This application of stochastic evolution is similar to that of Kandori (2003) who examines a public good game (without punishment option). Kandori, in line with psychological game theory, assumes intrinsic motivation to adhere to norms as long as others adhere to it and analyzes the resulting coordination problem. Occasional mutations — deviations from best responses — cause shifts of the norm. Downward shifts require fewer mutations than upward shifts in Kandori's model. As a result, high contribution levels eventually decay and only low contributions prevail in the long run, exactly as observed in experiments (see Ledyard, 1995). As shown below, adding a punishment option to the public good game reverses Kandori's result despite the fact that the commitment to punishment is limited. Small upward shifts of norms require fewer mutations than any downward shifts under a simple punishment rule. Therefore, for a low rate of mutations, shifts, conditional that they happen, are almost always upward and the stochastic evolution converges to high contribution levels. The evolution can be observed in the laboratory also in this case: the contribution level typically increases during public good experiments with punishment option (Fehr and Gächter, 2000).

The paper at hand does *not* examine where the authority's limited commitment ability comes from. For that reason, I choose a black box approach for the motivation of punishment. The authority is assumed to be able to commit to limited punishment.

There is experimental evidence supporting this assumption for the case of peer enforcement (e.g. Fehr and Gächter, 2000, 2002; Yamagishi, 1986). Punishment is modelled in this paper as an automatic, limited reaction governed by a punishment rule which is a function of the individual compliance levels. The focus *is* on specifying rules assuring high compliance under the constraint of limited punishment.

The analysis starts by examining an optimal punishment rule suitable for a collusion-resistant authority in section 2. A collusion-vulnerable authority and its associated coordination problem is studied in section 3. Section 4 concludes.

2 Punishment Rule Suitable for a Limited Collusion-Resistant Authority

This section reproduces the model in Steiner (2005). It formalizes the introductory argument that a collusion-resistant authority can always avoid non-compliance equilibria. Though the authority of this section could be centralized or decentralized, the model is formulated in the former setting, as I discuss its connection to tax enforcement at the end of the section.

Each player $i \in \mathcal{I} = \{1, \ldots, N\}$, $N \geq 1$, simultaneously chooses an action c_i from a common strategy set $S = \{0, \Delta, 2\Delta, \ldots, L\Delta\}$, where Δ is sufficiently small, $\Delta < 1$, and $L\Delta \geq N$. The assumption of the dense grid is needed to enable a sufficiently small increase in compliance. The grid is used as a technically convenient approximation of the continuous strategy space, so the assumption is not substantial. The assumption $L\Delta \geq N$ assures that players are not physically precluded from high compliance. The action profile of all players is denoted by $\mathbf{c}$.

The authority has committed to a punishment rule $\mathbf{p}(.)$, $\mathbf{p} : S^N \to \mathbb{R}^N_+$ that allocates punishment $p_i(\mathbf{c}) \geq 0$ to each player i after the authority observes the realized strategy profile. The authority committed to the rule before the players choose actions and the commitment has been commonly observed by all players. The payoffs of the players are

$$u_i(\mathbf{c}) = -c_i - p_i(\mathbf{c}). \tag{1}$$

Thus c^i is interpreted as the cost of compliance net of individual benefits of the compliance, if these exist. $(\mathcal{I}, S^N, \{u^i\}_{i=1}^N)$ is the *punishment game*. Only the one-stage interaction of players is modelled here; the behavior of the (limited) authority is an

assumption.

Enforcement of high compliance would be trivial if the authority could commit to any punishment rule. However, the authority is limited in the sense that it is at most able to commit to spending on punishment one unit per agent:

A1: $\frac{\sum_{i=1}^{N} p_i(\mathbf{c})}{N} \leq 1$ for any $\mathbf{c}$.

Despite assumption **A1**, there exists a punishment rule that induces a game with a unique equilibrium in which the actions of all players are approximately N. Denote the highest level below N by m_{cen}, the lowest action among players by l, and the second lowest by s with the convention that $l = s$ if there is more than one player with the lowest action. Let the punishment rule be

$$p_i(\mathbf{c}) = \begin{cases} \frac{N}{m_{cen}}\Big(\min(s, m_{cen}) - c_i\Big) & \text{if } c_i = l,\ l < s, \text{ and } c_i < m_{cen}, \\ 1 & \text{if } c_i = l,\ l = s, \text{ and } c_i < m_{cen}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The marginal punishment, which is $\frac{N}{m_{cen}} > 1$ or $\frac{1}{\Delta} > 1$, suffices to motivate the player with the lowest action to increase her action, as long as the lowest action is below m_{cen}. Yet the total punishment expenditures are always at most N because the punishment is not too costly even in situations when many players coordinate on the same lowest level, as then $s = l$ and each colluder is punished only slightly. This exact punishment rule is not necessarily practiced in reality; Proposition 1 simply demonstrates that a rule inducing high compliance exists.

Proposition 1. *1. The punishment game with punishment rule (2) has a unique equilibrium with all N players playing m_{cen}.*

2. Punishment rule (2) satisfies assumption **A1**.

Proof of the Proposition 1. **1.** Actions larger than m_{cen} are dominated by m_{cen} because a player who has chosen at least m_{cen} is never punished. Moreover, the player with the lowest action below m_{cen} always wishes to increase her action by at least Δ because the increase of her compliance by Δ decreases her punishment by $\frac{N}{m_{cen}}\Delta > \Delta$ or by $1 > \Delta$. Hence, the lowest level, 0, is dominated by level Δ. After elimination of $\{0, \Delta, \ldots, k\Delta\}$, level $(k+1)\Delta$ is dominated by $(k+2)\Delta$ because $(k+1)\Delta$ would be the lowest action among the non-eliminated strategies, for $k = 0, \ldots, \frac{m_{cen}}{\Delta} - 2$. Thus, the game can be solved by iterated elimination of dominated strategies. Only m_{cen} survives this process.

2. There is either only one player with the lowest action, in which case she is the only one being punished. The punishment is largest in this case if $s = m_{cen}$ and $c_i = 0$.

Then the punishment is $\frac{N}{m_{cen}} m_{cen} = N$. Or there may be many players with the lowest action, in which case $s = l$ and each punishment is 1. Thus the cost is at most 1 unit per player in both cases. □

A limited authority fulfilling **A1** cannot enforce higher actions than N, as this is the highest possible punishment it can inflict on a deviator. The "punish-the-worst" rule is thus the optimal rule.

Alm and McKee (2004) experimentally study several tax enforcement schemes and document that a rule similar to the "punish-the-worst" rule indeed elicits high compliance. The authors assume a coordination problem analogous to the one in the present model: audit probability increases with the difference between the average and agent's reported income. This models the use of the Discriminant Index Function (DIF) scores by the Internal Revenue Service in the United States. DIF is a statistical score indicating levels of suspiciousness of tax returns; those with above average DIF are more likely to be audited. Such an endogenous audit probability rule leads to a coordination game, in which full evasion by all agents constitutes an equilibrium. The experiment demonstrates that adding a small probability of a randomly allocated audit in a case of perfect collusion prevents coordination on full evasion. The intuition is the same as in the model of this section. Indeed, the experimental data show a gradual increase in compliance, as players try to escape the gradually increasing lowest position.

3 Punishment Rule Suitable for a Limited Collusion-Vulnerable Authority

This section examines long run sustainable compliance levels under a collusion-vulnerable authority. Unlike in the previous section, such an authority cannot assure high compliance in the short or medium run because zero compliance always constitutes an equilibrium. To compare the effectiveness of different punishment rules, I assume that players occasionally, but rarely, experiment with a randomly chosen action. I look for compliance levels that prevail in the long run.

For the sake of concreteness, the model is formulated in the setting of a public good game with punishment option which mimics in gross features the experiments in Fehr and Gächter (2000, 2002). The next subsection describes the evolution in a fixed group of players. A modification describing the evolution under a random matching protocol is given in subsection 3.2.

3.1 Partners Treatment

A fixed set of $N \geq 3$ risk-neutral players repeatedly plays the public good game with punishment option in rounds $t \in \mathbb{N}$, and each player i chooses a contribution level c_i^t from the common strategy set $S = \{0, \Delta, 2\Delta, \ldots, L\Delta\}$, $L\Delta \geq N$. S is of the same structure as in section 2 but a denser grid is required, $\Delta < \frac{1}{N-1}$. After the contributions $\mathbf{c}$ of all players are made and observed by everyone, players automatically assign punishment points to each other; p_j^i denotes the punishment i assigns to j.

The punishment $p_j^i(\mathbf{c}^{t-1}, \mathbf{c}^t)$ depends on the contribution levels of the previous and current rounds in this section; $p_j^i : S^N \times S^N \to \mathbb{R}_+$. By allowing mild history dependence, the model diverges from the experimental design of the partners treatment in Fehr and Gächter (2000), who excluded it in order to avoid reputation effects. The reputation effects are excluded here by assuming myopic behavior. I can therefore permit history-dependent punishment rules which are psychologically plausible and which allow higher contributions than do memoryless rules. Although longer memories could be considered, memory of length one turns out to be sufficient to support contribution levels of approximately N, which is the highest possible level. History dependence in not substantial for the qualitative results of the model. The enforceable contribution level increases linearly in the number of players even under a memoryless rule, but as $\sim \frac{N}{2}$ instead of $\sim N$. Only memoryless punishment rules are considered under the random matching setup in subsection 3.2.

Players play myopic best responses to the previous action profile in each round t.[2] That is, each player maximizes payoff under the punishment rules assuming that her opponents will carry over their contributions $\mathbf{c}^{t-1}$ from the last round:

$$c_i^t \in \arg\max_{c_i} \left\{ -c_i - \sum_{j \neq i} p_i^j \left(\mathbf{c}^{t-1}, \left(c_i, c_{-i}^{t-1} \right) \right) \right\}. \tag{3}$$

The public good does not enter the maximization problem; c_i is interpreted as the contribution costs net of the marginal increase of the public good. Also, the cost of the punishment does not enter the maximization problem although the agents bear the cost. The limited punishment is automatic and thus is not part of the agents' decision problem. Alternatively, I could presuppose a behavioral utility function under which the limited punishment would be optimal, but the main claim is that a small willingness to

[2]The results would not be changed if players could adjust to their best responses only with a certain probability.

punish leads to high contributions. The exact motivation to punish is outside the focus of this paper. The optimization problem (3) can be understood as a reduced form of a more complex optimization with the punishment stage already solved.

The strategy set S and the punishment rules $p_j^i(.,.)$ define a Markov process $(S^N, \mathbf{Q})$ where the transition matrix $\mathbf{Q}$ is determined by (3). Note that it is a memoryless process, despite the fact that the punishment rule is history dependent, because the optimization problem (3) depends only on the last round contribution profile $\mathbf{c}^{t-1}$. The pair $(S^N, \mathbf{Q})$ is the *unperturbed process*.

Assumption **A1** reformulated for the decentralized authority setting is:
A1': $\sum_{j \neq i} p_j^i \leq 1$ for all i and any $\mathbf{c}^{t-1}, \mathbf{c}^t$.
Assumption **A1'** is stricter than **A1** because it not only requires average expenses for the punishment to be below 1, but also individual expenses of each player to be below 1. The next assumption prohibits players from punishing peers that have contributed the same amount as themselves[3]:
A2: If $c_i^t = c_j^t$ then $p_j^i = 0$.

Assumption **A2** implies that $\mathbf{c} = \mathbf{0}$ is inevitably a steady state of the unperturbed process, so at worst a punishment rule does not induce any cooperation and at best there are multiple steady states. However, as demonstrated below, there exists a punishment rule under which increases of norms are much less demanding than decreases. Hence high contributions prevail in the long run.

In order to study the transitions between different steady states I introduce, following the framework of stochastic evolution of Kandori, Mailath and Rob (1993), occasional deviations from the unperturbed process: each player plays best response with probability $(1 - \epsilon)$ whereas with probability ϵ a "mutation" happens — the player chooses a random action from the uniform distribution on S. A *perturbed system* is a pair $(S^N, Q(\epsilon))$, where $Q(\epsilon)$ are the transition probabilities, with $\epsilon > 0$. The perturbed system has a unique invariant distribution μ^ϵ, which is close to $\mu^* \equiv \lim_{\epsilon \to 0} \mu^\epsilon$ for small ϵ. Ellison (2000) provides an intuitive "mutation counting" technique for the computation of μ^* based on the observation that step-by-step evolution passing through several intermediate states, with each step requiring few mutations, is quicker than a sudden evolutionary jump requiring the simultaneity of many mutations.

I utilize Ellison's observation and design a punishment rule under which only one mutation is needed for an increase in contributions by one level, but a decrease by any number of levels requires more than one mutation. As a consequence, evolution reaches

[3]Which implies that players never punish themselves.

high contribution levels more quickly than it escapes it. This intuition is formally expressed in the following proposition. Let m_{par} be the highest contribution level below $N-2$ and denote by M_{par} the state in which all players contribute m_{par}.

Proposition 2. *There exists a punishment rule satisfying* **A1'**, **A2** *under which M_{par} is the unique stochastically stable state, and the expected waiting time to reach M_{par} is of order $O(\epsilon^{-1})$.*

Proof of Proposition 2. The proof is based on the following lemma and the theorem in Ellison (2000).

Lemma 1. *There exists a punishment rule satisfying* **A1'**, **A2** *for which:*

1. Any common contribution level $0 \leq \overline{c} \leq m_{par}$, $\overline{c} \in S$ constitutes a steady state of the unperturbed process.

1'. No other limit sets of the unperturbed process than those in 1. exist.

2. Deviation of only one player from a steady state with common contribution level $\overline{c}$ suffices to induce transition to the steady state with level $\overline{c}+\Delta$, for any $\overline{c} < m_{par}$, $\overline{c} \in S$.

3. Deviation of more than one player from a steady state with common contribution level $\overline{c}$ is needed to induce transition to a steady state with a lower level, for any $\overline{c} \leq m_{par}$, $\overline{c} \in S$.

Proof of Lemma 1 is given Appendix A.

Having established Lemma 1, Proposition 2 is a consequence of Ellison's (2000) theorem that specifies the long run stochastically stable limit set in terms of radius and modified coradius. The radius $R(\Omega)$ is the number of mutations needed to escape Ω and hence property 3 in Lemma 1 and the fact that M_{par} is the highest steady state assures that $R(M_{par}) > 1$. The modified coradius $CR^*(\Omega)$ is the maximal modified number of mutations needed to reach Ω from other limit sets of the unperturbed process, where the modified number reflects that step-by-step evolution is more probable than sudden changes. In particular, a set Ω that is possible to reach through a series of one or zero mutation steps from anywhere has $CR^*(\Omega) = 1$; see Ellison (2000) for details. Property 2 in Lemma 1 guarantees that only one mutation is needed for transition from a steady state with level $\overline{c}$ to level $\overline{c}+\Delta$ and thus there is a path consisting of at most one mutation steps to M_{par} from any other state, and hence $CR^*(M_{par}) = 1$. According to theorem 2 in Ellison (2000), $R(M_{par}) > CR^*(M_{par})$ implies that M_{par} is the unique stochastically stable state. The same theorem specifies the waiting time as $O\left(\epsilon^{-CR^*(M_{par})}\right)$. □

Ellison provides an intuition for the speed of step-by-step evolution that translates naturally to the current setting: An increase in the norm by one contribution level is an ϵ probability event as it can be induced by one mutation. In contrast, a decrease in contribution level is an ϵ^2 or rarer event as at least two mutations are needed. Hence, conditional on a transition occurring, it is almost always an upward shift, for small ϵ.

It is worth noting that the waiting time $O(\epsilon^{-1})$ to reach M_{par} is of the least possible order. The contribution level enforceable by an authority limited by **A1'** and **A2** is bounded by $N-1$ because this is the maximal punishment a single deviator may suffer; thus the modified "punish-the-worst" rule induces a nearly optimal contribution level.

3.2 Strangers Treatment

The model of the partners treatment in the previous subsection describes evolution among a fixed set of players, evolving in isolation from the rest of the population. Alternatively, players may interact with different peers every round, in which case evolution occurs simultaneously in a large population, from which the groups are drawn anew each round. This subsection sketches evolution under the strangers treatment.

A population of KN risk-neutral players is randomly matched each round into $K \geq 2$ groups of $N \geq 2$ players to play the public good game with punishment option. The strategy set $S = \{0, \Delta, 2\Delta, \ldots, L\Delta\}$, $L\Delta \geq N$, is of the same structure as in sections 2 and 3.1 but the grid is denser, $\Delta < \frac{1}{KN-1}$. In each round, players can punish only the peers within the group they have been matched to and the punishment rules $p_j^i(\mathbf{c})$ are history *independent*, $p_j^i : S^N \to \mathbb{R}_+$. As in section 3.1, punishment rules are required to satisfy **A1'** and **A2**. The unperturbed process is again the best response dynamics and under the perturbed process, players choose the best response with probability $1-\epsilon$ and with probability ϵ choose a random action from the uniform distribution on S, as in section 3.1. Let m_{str} be the highest level below $(N-1)\frac{(K-1)N}{KN-1}$; it approaches $N-1$ for large K and N. Let M_{str} be the Markov state in which all players contribute m_{str}. The counterpart of Proposition 2 of subsection 3.1 is:

Proposition 3. *There exists a punishment rule satisfying* **A1'**, **A2**, *under which M_{str} is the unique stochastically stable state, and the expected waiting time to reach M_{str} is of order $O(\epsilon^{-1})$.*

Proof of the Proposition 3. The punishment rule (4 in Appendix A) without exception satisfies all four properties in Lemma 1.[4] Proof of property 1 and 1' remains unchanged.

[4] m_{par} needs to be replaced by m_{str} in the punishment rule and in Lemma 1.

Property 2 is implied by the inequality $\Delta < \frac{1}{KN-1}$: suppose there is a single deviator j contributing more than the norm $\overline{c}$ prescribes. Then the probability that j will be matched with i is $\frac{N-1}{KN-1}$, and hence i's expected punishment is $\frac{1}{N-1}\frac{N-1}{KN-1}$, which equals the right hand side of the inequality. Hence the inequality assures that one deviator is sufficient to induce all other players to increase their contributions by Δ.

The inequality $m_{str} < (N-1)\frac{(K-1)N}{KN-1}$ implies property 3: suppose that $c_k = \overline{c} \leq m_{par}$ for all $k \notin \{i, j\}$ and $c_j < \overline{c}$. Then a conservative estimate of the slope of the expected punishment for player i is $\frac{N-1}{m_{str}}\frac{(K-1)N}{KN-1} > 1$ because $\frac{(K-1)N}{KN-1}$ is the probability that j will not be in i's group, thus i will be the only deviator in her group, and hence punished by $\frac{N-1}{m_{str}}(\overline{c} - c_i)$.

The properties of Lemma 1 imply $R(M_{str}) > 1$, $CR^*(M_{str}) = 1$ and Proposition 3 is a consequence of Ellison's (2000) theorem as it was in Proposition 2 of subsection 3.1. □

The models in this section are not literal models of Fehr and Gächter's (2000, 2002) experiments. Their grids of contribution levels in the strangers treatment experiments were not as dense as Proposition 3 requires, the information structure of the partners treatment in the (2000) experiment precluded history-dependent punishment, and, on the other hand, punishment was cheaper in the experiments than in the model. Also, while experimental subjects may have had a variety of motivations for contributing, the model focuses solely on the contributions enforced by the threats of punishment. A combination of Kandori's (2003) model of intrinsic motivation and the models at hand could provide even higher estimates of sustainable contribution equilibrium than do the present models alone.

The models suggest that the high contributions are due to the game's structure that is focusing the limited committed resources of all players on one potential deviator. Keeping the commitment ability fixed, the contributions increase linearly with the number of players. This insight is experimentally confirmed by Carpenter (2005). who documents positive group size effects in public good games with punishment option even after controlling for the marginal group return of contributions.

Of course, the game requires quite a bit of information: the actions of all players need to be monitored, which is feasible in small groups such as work teams. Still, the effect can be noteworthy for a reasonable group size. Ten agents, each willing to spend only one unit for punishment, are able to collect at least $(10 - 2) \cdot 10 = 80$ units for a public good.

4 Conclusions

The models demonstrate that the commitment necessary for successful norms enforcement is small compared to the total cost of compliance of all agents. Agents in the compliance equilibrium consider deviating off the equilibrium *individually*. Hence, to support the compliance equilibrium, the authority needs only to be capable of substantially punishing one agent.

Nevertheless, other, non-compliance equilibria may exist. The main claim of the paper is that authorities can avoid these non-compliance equilibria by a proper punishment rule, even if their commitment capabilities are low. A punishment rule focusing on punishment of the worst offender creates competition among the agents and leads to a unique equilibrium with high compliance levels.

However, authorities using such a rule need to be able to punish perfectly colluding violators at least by a small amount, and many authorities fail to do so. Yet even such collusion-vulnerable authorities can avoid the non-compliance equilibria in the long run. They can introduce a punishment rule which deters revolts of a small fraction of players and enables a small fraction of players to initiate at least a tiny increase in compliance. Then, given a sufficiently small mutation rate, the increases are arbitrarily more times probable. High compliance prevails in the long run.

The prime application of the collusion-vulnerable authority model is the public good game with anger-driven punishment of free-riders. Even if the anger — a deviation from the *homo oeconomicus* framework — is limited, it can go a long way towards modifying equilibrium behavior. The public good game with punishment option is an instance of an institution that efficiently utilizes this behavioral deviation; a systematic search for other such institutions is needed.

References

[1] Alm J. and M. McKee, 2004, Tax Compliance as a Coordination Game, Journal of Economic Behavior & Organization 54, 297–312.

[2] Carpenter J., 2005, Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods, Games and Economic Behavior, forthcoming.

[3] Ellison G., 2000, Basins of Attraction, Long Run Stochastic Stability, and the Speed of Step-by-Step Evolution, Review of Economic Studies 67, 17–45.

[4] Fehr E. and S. Gächter, 2000, Cooperation and Punishment in Public Goods Experiments, American Economic Review 90, 980–994.

[5] Fehr E. and S. Gächter, 2002, Altruistic Punishment in Humans, Nature 415, 137–140.

[6] Kandori M., 2003, The Erosion and Sustainability of Norms and Morale, The Japanese Economic Review 54, 29–48.

[7] Kandori M., G. Mailath and R. Rob, 1993, Learning, Mutation, and Long Run Equilibria in Games, Econometrica 61, 29–56.

[8] Ledyard J., 1995, Public Goods: A Survey of Experimental Research, in: Kagel, J. and A. Roth, Eds, The Handbook of Experimental Economics, (Princeton University Press, Princeton) 111-194.

[9] Steiner J., 2005, A Trace of Anger is Enough, on the Enforcement of Social Norms, CERGE-EI Working paper No. 246.

[10] Yamagishi T., 1986, The Provision of a Sanctioning System as a Public Good, Journal of Personality and Social Psychology 51, 110–116.

A Proof of Lemma 1

Lemma 1. *There exists a punishment rule satisfying* **A1'**, **A2** *for which:*

1. Any common contribution level $0 \leq \overline{c} \leq m_{par}$, $\overline{c} \in S$ *constitutes a steady state of the unperturbed process.*

1'. No other limit sets of the unperturbed process than those in 1. exist.

2. Deviation of only one player from a steady state with common contribution level $\overline{c}$ *suffices to induce transition to the steady state with level* $\overline{c} + \Delta$*, for any* $\overline{c} < m_{par}$, $\overline{c} \in S$.

3. Deviation of more than one player from a steady state with common contribution level $\overline{c}$ *is needed to induce transition to a steady state with a lower level, for any* $\overline{c} \leq m_{par}$, $\overline{c} \in S$.

Proof of Lemma 1. let the definitions of l and s remain as in section 2. Consider a "modified punish-the-worst" rule:

$$p_j^i = \begin{cases} \frac{1}{m_{par}}\Big(\min(s, m_{par}) - c_j\Big) & \text{if } c_j = l,\ c_j < m_{par}, l < s, \text{ and } c_i > c_j, \\ \frac{1}{N-1} & \text{if } c_j = l,\ c_j < m_{par}, l = s, \text{ and } c_i > c_j, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

except for situations when player k starts a rebellion against a norm of a common contribution level $\overline{c}$ in the previous round $t-1$, persists in that rebellion in round t, and player j joins the rebellion; then the remaining players concentrate on punishing the new free-rider j, not the old k. Formally, the exception states that (4) does not apply at t when at $t-1$ all players $i \neq k$ contributed some common level $c_i^{t-1} = \overline{c} \leq m_{par}$ and k contributed $c_k^{t-1} < \overline{c}$, and $c_k^t = c_k^{t-1}$, $c_j^t < \overline{c}$, and $c_i^t = \overline{c}$, for $i \notin \{j,k\}$ in round t. Then all $N-2$ players $i \notin \{j,k\}$ punish j in round t each by amount

$$p_j^i = \frac{1}{m_{par}}(\overline{c} - c_j^t).$$

The punishment rule suitable for the strangers treatment does not employ this exception.

This rule satisfies **A1'** because either player punishes only one of her peers in which case she spends at most $\frac{1}{m_{par}} m_{par} = 1$ or she punishes many players and then she spends at most $(N-1)\frac{1}{N-1} = 1$. The rule satisfies **A2** because it prescribes to punish only peers who have contributed less than the punisher. Let us verify that the modified "punish-the-worst" rule satisfies all four properties in Lemma 1:

1. The best response to $\mathbf{c}^{t-1} = \overline{\mathbf{c}}$ and $c_{-i}^t = \overline{c}$ is[5] $c_i^{t*} = \overline{c}$. Hence a state in which all players contribute $\overline{c}$ is a steady state of the unperturbed process.

2. Suppose $c_i^t > \overline{c}$, $c_j^t = \overline{c} < m_{par}$ for all $j \neq i$. The best response of $j \neq i$ is $\overline{c} + \Delta$, the best response of i is $\overline{c}$. Thus, at $t+1$, $c_i^{t+1} = \overline{c}$, $c_j^{t+1} = \overline{c} + \Delta$, and at $t+2$ all players contribute $\overline{c} + \Delta$ which becomes the new steady state of the unperturbed process.

3. Suppose that only one player has deviated from the common contribution level in round t; $c_j^t = \overline{c} \leq m_{par}$ for all $j \neq i$ and $c_i^t < \overline{c}$. Then the exemption applies in $t+1$ and the best response of all players in $t+1$ is to contribute $\overline{c}$.

1'. Consider a state $\mathbf{c}$ in which more than one contribution level is chosen. Let us distinguish two cases: in case **A**, $N-1$ players contribute some common $\overline{c}$ and the contribution of only one player differs from $\overline{c}$; case **B** includes all other situations. If **A** arises, players converge to a common contribution level $\overline{c}$ or $\overline{c} + \Delta$ within one or two rounds(see proofs of properties 2 and 3). In case **B**, the best response of each player i is to contribute $l_{-i} + \Delta > l$, where l_{-i} is the lowest contribution among i's opponents. Therefore the lowest contribution increases in those rounds when case **B** arises.[6][7] Thus in each round either l increases or **A** arises, and because the set of

[5] To avoid confusion, $c_i^{t-1} = \overline{c}$ for all i and $c_j^t = \overline{c}$ for all $j \neq i$.

[6] This does not hold in situations described in the proof of property 2. Therefore the division of all situations into categories **A** and **B** is necessary.

[7] In the case of the adaptation of the proof for subsection 3.2 the best response is $l_{-i} + \Delta$ or higher.

the contribution levels is finite, either l converges to m_{par} or **A** arises and under both eventualities players converge to a common contribution level. □